Writing Domain Names in Different Arabic Script Based Languages

Language VS. Script

ICANN Regional Meeting, Dubai April 1-3 2008

Dr. Abdulaziz H. Al-Zoman
Director of SaudiNIC - CITC
Chairman of Steering Committee
Arabic Domain Name Pilot Project

Raed I. Al-Fayez
SaudiNIC - CITC
Chairman of Technical Committee
Arabic Domain Name Pilot Project

www.arabic-domains.org
Part I: Arabic language
  – Characteristics
  – Basic linguistic issues
  – Recommendations
  – Accepted character set table

Part II: What ccTLD communities are saying?

Part III: Let us expand and look at the whole script
  – About the Arabic script
  – Unicode character code chart
  – The Unicode codepoints and IDNA
  – Issues need further investigations

Part IV: Conclusions
• Arab Team for Arabic Domain Names
  – Formed under the auspices of Arab League, 2005
  – Supervising “ADNPP: Arabic Domain Name Pilot Project”

• SaudiNIC
  – Administering the Saudi domain name space (.sa) since 1995.
  – Operated by Communication and Information Technology Commission (CITC) … governmental org.
  – Leading the local community effort towards supporting Arabic language in DNS
    • Head of the ADNPP Steering Committee
    • Head of the ADNPP Technical Committee
Part I: Arabic language

www.arabic-domains.org
1. Written from right to left.

2. Most characters have different shapes
   - depending on their position (beginning, middle, or end) within a word
   - probably conjugated with preceding and succeeding characters.
   - These different shapes for a single character do not count as different code points but they are handled using different fonts
   - Letters that can be joined are always joined in both hand-written and printed Arabic (cursive style).
3. Two sets of **numerals** are used:
   - Arabic: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
   - Arabic-Indic: ٠١٢٣٤٥٦٧٨٩
   - Numerals are written from **left to right**

4. Tashkeel (diacritic)
   - A small sing (not a letter) that is usually put on top or under a character for the purpose of correct pronunciation
   - It is **not widely used** except incase of the possibility of mispronouncing words that have the same letters but with different pronunciations, and hence having different meanings.

5. Abbreviations are **not widely used**
   - When an abbreviation is written (in domain name) characters will be joined together ... leads to a different word and pronunciation
Identified a number of linguistic issues with respect to domain names:

- Usage of Tashkeel (diacritics).
- Usage of Kashidah.
- Character folding.
- Which numerical digits should be supported.
- Connecting multiple words.
- Mixing with other languages.
- Usage of special characters.
Basic Linguistic Issues
Tashkeel (Diacritics)

- Not a letter but a small sing that is usually put on top or under an Arabic letter for the purpose of correct pronunciation.
- It is not widely used except in case of the possibility of mispronouncing words that have the same letters but with different pronunciations, and hence having different meanings.
- If supported, there will be huge number of combinations when registering a single domain name.
- Good playground for phishing.

Recommendation:
Tashkeel can be supported only in the user interface but should not be stored in the zone file. Therefore, it can be striped off at the preparation of internationalized strings ("stringprep") phase.

www.arabic-domains.org
Not a letter. It is a horizontal line (like dash) used to lengthen the connection line between letters.

Used sometimes to enhance the display of Arabic words on screens or printouts.

Good playground for domain name phishing

Basic Linguistic Issues
Kasheeda (Tatweel)

Recommendation:
Kasheeda should not be used in Arabic domain names
A character folding is the process where multiple letters (that may have some similarity with respect to their shapes) are folded into one shape. This includes:
- Folding Teh Marbuta and Heh at the end of a word.
- Folding different forms of Hamzah.
- Folding Alif Maksura and Yeh at the end of a word.
- Folding Waw with Hamzah and Waw.

Character folding is unacceptable in the Arabic language because it changes the meaning of the words and it is against the simplest spelling rules. Replacing a character with another character, which may have the same shape but different pronunciation, will give a different meaning. This will lead to have only one form (word) out many other forms of words that are made by all the combination of folded characters. Hence, the other forms will be masked by the common form.

Hand writing mixes between different characters (e.g., Heh and Teh-Marbuta) and this is due to laziness or weakness in spelling. However, this is not the case in published and printed materials. One of the motivations to support the Arabic language in domain names is to preserve the language particularly with the spread of the globalization movement. Hence, character folding is working against this motivation since it is going to have a negative effect on the principles and ethics of the language.

Recommendation:
Character folding should not be allowed
Basic Linguistic Issues

Numbers

- In the Arab world, there are two sets of numerical digits used:
  - Set I: (0, 1, 2, 3, 4, 5, 6, 7, 8, 9),
    - Mostly used in the western part of the Arab world (al-maghrib al-arabi).
  - Set II: (۰، ۱، ۲، ۳، ۴، ۵، ۶، ۷، ۸)،
    - Mostly used in the eastern part of the Arab world (al-mashriq al-arabi).

- Their writing direction (LTR) differs than the writing direction in the Arabic letters (RTL)

Recommendation:
Both sets should be supported in the user interface and both are folded to one set (Set I) at the preparation of internationalized strings (e.g., "stringprep") phase.
Basic Linguistic Issues
Connecting Multiple Words

- In the Arab language words are separated by spaces.
- Connecting words without spaces is usually not acceptable.
  - Decrease readability

Recommendation:
Space should be used to separate words if it is technically visible. Otherwise, it is recommended that multiple words are separated by the character "-" dash.
If the space is used as a word separator in Arabic domain names then it should be only a single space and it should not be used at the beginning or at the end of words.
### Arabic Linguistic Issues

#### Recommendation Summary

<table>
<thead>
<tr>
<th>Tashkeel (Diacritics)</th>
<th>Tashkeel should not be allowed. However, if there is a need to allow users to enter it as part of a domain name then it should be stripped off by nameprep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kasheeda</td>
<td>Kasheeda should be disallowed</td>
</tr>
<tr>
<td>Character folding:</td>
<td></td>
</tr>
<tr>
<td>Teh Marbuta + Heh</td>
<td>Folding should not be allowed</td>
</tr>
<tr>
<td>different forms of Hamzah</td>
<td></td>
</tr>
<tr>
<td>Alif Maqura+YaNumbers</td>
<td></td>
</tr>
<tr>
<td>Numbers (numerical digits)</td>
<td>If it is technically possible, it is preferred to support both (Latin and Arabic) sets with folding to one set. Otherwise, Latin set is sufficient</td>
</tr>
<tr>
<td>Connecting Multiple Words</td>
<td>It is recommended that multiple words are separated by the character &quot;-&quot;.</td>
</tr>
<tr>
<td>Mixing Latin and Arabic Characters</td>
<td>It is recommended that Arabic domain names be pure Arabic and they should not be mixed with other languages.</td>
</tr>
<tr>
<td>Special Characters (e.g., @, #, $, %, ...)</td>
<td>It is recommended that Arabic domain names should follow the standard with respect to the use of special characters.</td>
</tr>
</tbody>
</table>
• Development Principles
  – Study and discuss the linguistic issues
    • extensively (more than 3 years) by a team who work under the umbrella of Arab League
  – Adhere to LDH convention
  – Follow the inclusion mode recommended by the new IDNA standards
    • Many characters are disallowed: symbols, punctuations, diacritics (tashkeel), Koranic annotation signs, honorifics, etc are not allowed.
  – Arabic labels are not intended to write sentences or phrases

• The goal is to develop:
  – A simple character set table that includes ONLY needed letters and digits
The following Unicode characters are accepted only:

<table>
<thead>
<tr>
<th>Unicode Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>U0621 - U063A</td>
<td>(hamza-gheen)</td>
</tr>
<tr>
<td>U0641 - U064A</td>
<td>(feh-yeh)</td>
</tr>
<tr>
<td>U0660 - U0669</td>
<td>(٠،۰۰۰۰۰۰۰۰)</td>
</tr>
<tr>
<td>U0030 - U0039</td>
<td>(0,1,2,...,9)</td>
</tr>
<tr>
<td>U002D</td>
<td>(Hyphen)</td>
</tr>
<tr>
<td>U002E</td>
<td>(Dot)</td>
</tr>
</tbody>
</table>

- It is extremely important to standardize this character set table for the Arabic Language.
### Arabic Language Accepted Character Set Table

#### Characters from Unicode Arabic Table (0600–06FF)

| \( \text{U+0621} \) (א) | Arabic Letter HAMZA | \( \text{U+0638} \) (מ) | Arabic Letter ZAH |
| \( \text{U+0622} \) (ל) | Arabic Letter ALEF with MADDA above | \( \text{U+0639} \) (ע) | Arabic Letter AIN |
| \( \text{U+0623} \) (מ) | Arabic Letter ALEF with HAMZA above | \( \text{U+063A} \) (ל) | Arabic Letter GHAIN |
| \( \text{U+0624} \) (ט) | Arabic Letter WAW with HAMZA above | \( \text{U+0641} \) (ך) | Arabic Letter FEH |
| \( \text{U+0625} \) (י) | Arabic Letter ALEF with HAMZA below | \( \text{U+0642} \) (כ) | Arabic Letter QAF |
| \( \text{U+0626} \) (גי) | Arabic Letter YEH with HAMZA above | \( \text{U+0643} \) (יך) | Arabic Letter KAF |
| \( \text{U+0627} \) (ך) | Arabic Letter ALEF | \( \text{U+0644} \) (ך) | Arabic Letter LAM |
| \( \text{U+0628} \) (ם) | Arabic Letter BEH | \( \text{U+0645} \) (ך) | Arabic Letter MEEM |
| \( \text{U+0629} \) (ת) | Arabic Letter TEH MARBUTA | \( \text{U+0646} \) (ך) | Arabic Letter NOON |
| \( \text{U+062A} \) (ת) | Arabic Letter TEH | \( \text{U+0647} \) (ך) | Arabic Letter HEH |
| \( \text{U+062B} \) (ף) | Arabic Letter THEH | \( \text{U+0648} \) (ך) | Arabic Letter WAW |
| \( \text{U+062C} \) (ץ) | Arabic Letter JEEM | \( \text{U+0649} \) (ך) | Arabic Letter ALEF MAKSURA |
| \( \text{U+062D} \) (ץ) | Arabic Letter HAH | \( \text{U+064A} \) (ך) | Arabic Letter YEH |
| \( \text{U+062E} \) (ץ) | Arabic Letter KHAH | \( \text{U+0660} \) (ך) | Arabic-Indic Digit Zero |
| \( \text{U+062F} \) (ם) | Arabic Letter DAL | \( \text{U+0661} \) (ך) | Arabic-Indic Digit One |
| \( \text{U+0630} \) (ס) | Arabic Letter THAL | \( \text{U+0662} \) (ך) | Arabic-Indic Digit Two |
| \( \text{U+0631} \) (ץ) | Arabic Letter REH | \( \text{U+0663} \) (ך) | Arabic-Indic Digit Three |
| \( \text{U+0632} \) (ץ) | Arabic Letter ZAIN | \( \text{U+0664} \) (ך) | Arabic-Indic Digit Four |
| \( \text{U+0633} \) (ך) | Arabic Letter SEEN | \( \text{U+0665} \) (ך) | Arabic-Indic Digit Five |
| \( \text{U+0634} \) (ך) | Arabic Letter SHEEN | \( \text{U+0666} \) (ך) | Arabic-Indic Digit Six |
| \( \text{U+0635} \) (ך) | Arabic Letter SAD | \( \text{U+0667} \) (ך) | Arabic-Indic Digit Seven |
| \( \text{U+0636} \) (ך) | Arabic Letter DAD | \( \text{U+0668} \) (ך) | Arabic-Indic Digit Eight |
| \( \text{U+0637} \) (ך) | Arabic Letter TAH | \( \text{U+0669} \) (ך) | Arabic-Indic Digit Nine |
- Characters from Unicode Basic Latin Table (0000–007F):
  - 0030 (0)  Digit Zero
  - 0031 (1)  Digit One
  - 0032 (2)  Digit Two
  - 0033 (3)  Digit Three
  - 0034 (4)  Digit Four
  - 0035 (5)  Digit Five
  - 0036 (6)  Digit Six
  - 0037 (7)  Digit Seven
  - 0038 (8)  Digit Eight
  - 0039 (9)  Digit Nine
  - 002D (-)  Hyphen-Minus
  - 002E (.)  Full Stop (Dot)
Part II: What ccTLD Communities are saying?

www.arabic-domains.org
What ccTLD Communities are saying?

- **2005**: Chinese Domain Name Consortium (CDNC)
  - The Proposal for Internationalizing ccTLD Names
    - Give the priority to internationalizing ccTLDs.
      - To ensure the system stability, it’s recommended to IDN ccTLD before IDN gTLD.
    - One form of language character variant of internationalized ccTLD is accepted.
    - ccTLD registries make their own choice of which IDN character sets.
    - Register and operate the internationalized ccTLDs in the root DNS server in the form of IDNA Punycode.
  - Supported by SaudiNIC

- **2007**: APTLD Position on Top Level Internationalised Domain Names
  - Allow each existing ccTLD to manage one additional country or territory specific ccTLD in a recognised non-ASCII script of their country or territory.
  - Support from CENTR
  - Support from AFTLD
What ccTLD Communities are saying?

... URLs

- Chinese Domain Name Consortium (CDNC)
  - The Proposal for Internationalizing ccTLD Names, June 2005
  - Supported by SaudiNIC

- APTLD Position on Top Level Internationalised Domain Names - June 21, 2007
  - Support from CENTR (www.centr.org)
  - Support from AFTLD (www.aftld.org)
For the purpose of speeding up the deployment of IDN.IDN (i.e., internationalized top level domains), the following principles are agreed on:

- **Give the priority to internationalizing ccTLDs.** To ensure the system stability, it is recommended to internationalize ccTLDs before internationalizing gTLDs.
- One IDN ccTLD string per one ASCII ccTLD should be applied by the current ASCII ccTLD registry and introduced to the root zone in the initial phase. Recognizing that some countries or regions may have character or language variants, initially only one form of character sets shall be chosen for IDN use by each sponsored registry. Relaxation of this limitation should be revisited after initial deployment of IDN ccTLDs has been experienced.
- The initial choice of IDN ccTLD is made by the current ccTLD Registry with ample comment period for possible objections and reservations by other relevant bodies.
- Register and operate the IDN ccTLDs in the root DNS server in the form of IDNA Punycode.
- With respect to languages using Arabic script, **it is recommended that each language be treated separately**, i.e., each language (e.g., Arabic, Persian, Urdu, ...) have its own accepted character set that should be applied for domain name registration for that language only.
- Supporting **multiple languages that use the same script in an IDN-TLD** (e.g., IDN gTLDs) demands further studies and investigations that should not prevent the implementation of IDN at the ccTLD level.
Part III:
Let us expand and look at the whole script

www.arabic-domains.org
References

- New IDNA200x documents
  - www.ietf.org
- PAN10
  - www.PANL10n.net
- Allowable Characters (under dot-iran)
  - www.nic.ir/Allowable_Characters_dot-iran
- Unicode
  - www.unicode.org
- ADNPP
  - www.arabic-domains.org
About Arabic Script

- The 2nd most widely used alphabetic writing system in the world
- Used by many languages such as:
  - Persian, Urdu, Turkish, Kurdish, Pashto, Swahili, ...
- It is widely used by more than 43 countries
  - more than one billion potential users could be concerned in using Arabic script domain names.
Arabic Script
Unicode Character Code Chart

- Arabic language character codes occupy a very small section of the Arabic script Unicode chart
- Arabic script Unicode chart is used by many other languages
  - that may add or change characters to represent phonemes that do not appear in Arabic phonology.
  - A new character usually created based on modifying the basic shape of an existing Arabic character, for example, by adding more dots.
  - These additions have meaning to the new language but not to the original language.
- Therefore, many characters would easily be confused with some other characters from other languages
- Some languages may *add or change characters* to represent phonemes that do not appear in Arabic phonology.
- A new character usually created based on *modifying the basic shape* of an existing Arabic character, for example, by adding more dots.
<table>
<thead>
<tr>
<th></th>
<th>060</th>
<th>061</th>
<th>062</th>
<th>063</th>
<th>064</th>
<th>065</th>
<th>066</th>
<th>067</th>
<th>068</th>
<th>069</th>
<th>06A</th>
<th>06B</th>
<th>06C</th>
<th>06D</th>
<th>06E</th>
<th>06F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Codepoints</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0600..0603</td>
<td>CONTEXTO # ARABIC NUMBER SIGN..ARABIC SIGN SAFHA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0604..060A</td>
<td>UNASSIGNED # &lt;reserved&gt;..&lt;reserved&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>060B..060F</td>
<td>DISALLOWED # AFGHANI SIGN..ARABIC SIGN MISRA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0610..0615</td>
<td>PVALID # ARABIC SIGN SALLALAHOU ALAYHE WASSALLAM..AR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0616..061A</td>
<td>UNASSIGNED # &lt;reserved&gt;..&lt;reserved&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>061B</td>
<td>DISALLOWED # ARABIC SEMICOLON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>061C..061D</td>
<td>UNASSIGNED # &lt;reserved&gt;..&lt;reserved&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>061E..061F</td>
<td>DISALLOWED # ARABIC TRIPLE DOT PUNCTUATION MARK..ARABIC Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0620</td>
<td>UNASSIGNED # &lt;reserved&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0621..063A</td>
<td>PVALID # ARABIC LETTER HAMZA..ARABIC LETTER GHAIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>063B..063F</td>
<td>UNASSIGNED # &lt;reserved&gt;..&lt;reserved&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0640..065E</td>
<td>PVALID # ARABIC TATWEEL..ARABIC FATHA WITH TWO DOTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>065F</td>
<td>UNASSIGNED # &lt;reserved&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0660..0669</td>
<td>PVALID # ARABIC-INDIC DIGIT ZERO..ARABIC-INDIC DIGIT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>066A..066D</td>
<td>DISALLOWED # ARABIC PERCENT SIGN..ARABIC FIVE POINTED STA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>066E..0674</td>
<td>PVALID # ARABIC LETTER DOTLESS BEH..ARABIC LETTER HIG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0675..0678</td>
<td>DISALLOWED # ARABIC LETTER HIGH HAMZA ALEF..ARABIC LETTER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0679..06D3</td>
<td>PVALID # ARABIC LETTER TTEH..ARABIC LETTER YEH BARREE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06D4</td>
<td>DISALLOWED # ARABIC FULL STOP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06D5..06DC</td>
<td>PVALID # ARABIC LETTER AE..ARABIC SMALL HIGH SEEN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06DD</td>
<td>CONTEXTO # ARABIC END OF AYAH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06DE</td>
<td>DISALLOWED # ARABIC START OF RUB EL HIZB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06DF..06E8</td>
<td>PVALID # ARABIC SMALL HIGH ROUNDED ZERO..ARABIC SMALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06E9</td>
<td>DISALLOWED # ARABIC PLACE OF SAJDAH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06EA..06FC</td>
<td>PVALID # ARABIC EMPTY CENTRE LOW STOP..ARABIC LETTER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06FD..06FE</td>
<td>DISALLOWED # ARABIC SIGN SINDHI AMPERSAND..ARABIC SIGN SI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06FF</td>
<td>PVALID # ARABIC LETTER HEH WITH INVERTED V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: draft-faltstrom-idnbis-tables-05.txt
... in general, characters are placed into "DISALLOWED" if they fall into one or more of the following groups:

- The character is a symbol or punctuation form or, more generally, something that is not a letter or digit.”

“Symbols, drawing characters, and various notational characters are permanently excluded …”

John Klensin, Internet-Draft: IDNA200X Rationale, Feb 2008
Undesirable Characters
Punctuations, Signs, Marks, Annotations, Honorifics

- Subtending marks 0600 - 0603
- Currency sign 060B
- Punctuations 060C-060D, 061B, 061E-061F, 066A-066D, 06D4
- Poetic marks 060E-060F
- Honorifics 0610-0614
- Koranic annotation sign 0615, 06D6-06ED
- Tatweel 0640
- Points 064B-0652, 0670
- The use of combining marks with some base characters would confuse with other character.
  - Combining maddah and hamza 0653-0655
  - Other combining marks 0656-065E

\[ \text{U+063} \neq \text{U+063} + \text{U+063} \]

U+0623 U+0654 U+0627
أنا ≠ 

أنا

输入：
input[0] = U+0627
input[1] = U+0654
input[2] = U+066e
input[3] = U+06ec
input[4] = U+0627

输入：
input[0] = U+0623
input[1] = U+0646
input[2] = U+0627
Issues Need Further Investigations

Diacritics (Tashkeel)

- Tashkeel Points: about 9
  - 064B-0652, 0670
- From user point view, he/she does not know
  - Whether the label is with or without tashkeel
  - Whether the label is totally (i.e., all characters) use tashkeel
  - The correct pronunciations to put the correct taskeel
  - There are many different pronunciations depending on the local dialogue
- There will be huge number of combinations when registering a single label
  - good playground for phishing
The support of **ZWJ** and **ZWNJ** in domain names would result in some **confusions**.

Their use and concepts are **not known** to regular users.
- ZWNJ: Visually noticed

input[0] = U+062d
input[1] = U+200c
input[2] = U+0628
input[3] = U+0644

input[0] = U+062d
input[1] = U+0628
input[2] = U+0644
- ZWNJ: Visually Unnoticed

\[ \text{input}[0] = \text{U+0637} \]
\[ \text{input}[1] = \text{U+200c} \]
\[ \text{input}[2] = \text{U+0628} \]
\[ \text{input}[3] = \text{U+0644} \]
Issues Need Further Investigations

ZWNJ & ZWJ Control Characters

ZWNJ & ZWJ Control Characters

ZWJ Visually noticed

ZWJ Visually Unnoticed
Issues Need Further Investigations

Arabic-Indic VS. Eastern Arabic-Indic digits

1٢٣٤٥٦٧٨٩٠

≠

1٢٣٤٥٦٧٨٩٠

input[0] = U+06f1
input[1] = U+06f2
input[2] = U+06f3
input[3] = U+06f7
input[4] = U+06f8
input[5] = U+06f9
input[6] = U+06f0

input[0] = U+0661
input[1] = U+0662
input[2] = U+0663
input[3] = U+0667
input[4] = U+0668
input[5] = U+0669
input[6] = U+0660
Issues Need Further Investigations

**Digits**

- Arabic-Indic vs. European-Arabic digits
  - ٠١٢٣٤٥٦٧٨٩
  - 0 1 2 3 4 5 6 7 8 9
- "Windows has supported number substitution by allowing the representation of different cultural shapes for the same digits while keeping the internal storage of these digits unified among different locales, for example numbers are stored in their well known hexadecimal values, 0x40, 0x41, but displayed according to the selected language."


```
<table>
<thead>
<tr>
<th>Input</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>input[0]</td>
<td>U+0645</td>
</tr>
<tr>
<td>input[1]</td>
<td>U+0031</td>
</tr>
<tr>
<td>input[2]</td>
<td>U+0032</td>
</tr>
<tr>
<td>input[3]</td>
<td>U+0645</td>
</tr>
</tbody>
</table>

≠

<table>
<thead>
<tr>
<th>Input</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>input[0]</td>
<td>U+0645</td>
</tr>
<tr>
<td>input[1]</td>
<td>U+0661</td>
</tr>
<tr>
<td>input[2]</td>
<td>U+0662</td>
</tr>
<tr>
<td>input[3]</td>
<td>U+0645</td>
</tr>
</tbody>
</table>
```
There are a number of groups of characters that have the same shapes,
- eg. Kaf, Heh, Yeh, Alef, … groups
<table>
<thead>
<tr>
<th>Char.</th>
<th>Code</th>
<th>Arabic Letter</th>
<th>Confusing Letters From Other Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>١</td>
<td>0621</td>
<td>Hamza</td>
<td>١ (0674), ١ (06FD)</td>
</tr>
<tr>
<td>١</td>
<td>0622</td>
<td>Alef With Madda Above</td>
<td>١ (0671)</td>
</tr>
<tr>
<td>١</td>
<td>0623</td>
<td>Alef With Hamza Above</td>
<td>١ (0671), ١ (0672), ١ (0675)</td>
</tr>
<tr>
<td>١</td>
<td>0624</td>
<td>Waw With Hamza Above</td>
<td>١ (0676), ١ (0677), ؤ (06C6), ٩ (06C8), ٩ (06C9), ٩ (06CF)</td>
</tr>
<tr>
<td>١</td>
<td>0625</td>
<td>Alef With Hamza Below</td>
<td>١ (0673)</td>
</tr>
<tr>
<td>١</td>
<td>0626</td>
<td>Yeh With Hamza Above</td>
<td>١ (0678), ١ (06CE), ١ (06D3)</td>
</tr>
<tr>
<td>١</td>
<td>0627</td>
<td>Alef</td>
<td></td>
</tr>
<tr>
<td>ب</td>
<td>0628</td>
<td>Beh</td>
<td>ب (066E), ث (0679), ث (067E)</td>
</tr>
<tr>
<td>ء</td>
<td>0629</td>
<td>The Marbuta</td>
<td>ء (06C3), ء (06C0), ء (06C2)</td>
</tr>
<tr>
<td>Char.</td>
<td>Code</td>
<td>Arabic Letter</td>
<td>Confusing Letters From Other Languages</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>--------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>ت</td>
<td>062A</td>
<td>Teh</td>
<td>ت (067A), ت (067C)</td>
</tr>
<tr>
<td>ث</td>
<td>062B</td>
<td>Thah</td>
<td>ث (067D), ث (067F)</td>
</tr>
<tr>
<td>ج</td>
<td>062C</td>
<td>Jeem</td>
<td>ج (0683), ج (0684)</td>
</tr>
<tr>
<td>ح</td>
<td>062D</td>
<td>Hahe</td>
<td>ح (0681)</td>
</tr>
<tr>
<td>خ</td>
<td>062E</td>
<td>Khah</td>
<td>خ (0681), خ (0682)</td>
</tr>
<tr>
<td>د</td>
<td>062F</td>
<td>Dal</td>
<td>د (0688), د (0689), د (068A), پ (068D)</td>
</tr>
<tr>
<td>ذ</td>
<td>0630</td>
<td>Thal</td>
<td>ذ (0688), ذ (068C)</td>
</tr>
<tr>
<td>ر</td>
<td>0631</td>
<td>Reh</td>
<td>ر (0691), ر (0693), ر (0695)</td>
</tr>
<tr>
<td>ز</td>
<td>0632</td>
<td>Zain</td>
<td>ز (0691), ز (0692), ز (0697)</td>
</tr>
<tr>
<td>Char.</td>
<td>Code</td>
<td>Arabic Letter</td>
<td>Confusing Letters From Other Languages</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>---------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>ﺔ</td>
<td>0633</td>
<td>Seen</td>
<td></td>
</tr>
<tr>
<td>ﺔ</td>
<td>0634</td>
<td>Sheen</td>
<td>ﺔ (06FA)</td>
</tr>
<tr>
<td>ﺔ</td>
<td>0635</td>
<td>Sad</td>
<td>ﺔ (069D)</td>
</tr>
<tr>
<td>ﺔ</td>
<td>0636</td>
<td>Dad</td>
<td>ﺔ (069E), ﺔ (06FB)</td>
</tr>
<tr>
<td>ﺔ</td>
<td>0637</td>
<td>Tah</td>
<td></td>
</tr>
<tr>
<td>ﺔ</td>
<td>0638</td>
<td>Zah</td>
<td>ﺔ (069F)</td>
</tr>
<tr>
<td>ﺔ</td>
<td>0639</td>
<td>Ain</td>
<td></td>
</tr>
<tr>
<td>ﺔ</td>
<td>063A</td>
<td>Ghain</td>
<td>ﺔ (06A0)</td>
</tr>
<tr>
<td>ﺔ</td>
<td>0641</td>
<td>Feh</td>
<td>ﺔ (06A1), ﺔ (06A2), ﺔ (06A3), ﺔ (06A4)</td>
</tr>
<tr>
<td>Char.</td>
<td>Code</td>
<td>Arabic Letter</td>
<td>Confusing Letters From Other Languages</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>---------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>ق</td>
<td>0642</td>
<td>Qaf</td>
<td>ق (066F), ق (06A8)</td>
</tr>
<tr>
<td>ك</td>
<td>0643</td>
<td>Kaf</td>
<td>ك (06A9), ك (06AA), ك (06AB), ك (06AC), ك (06AF), ك (06B2), ك (06B3)</td>
</tr>
<tr>
<td>ل</td>
<td>0644</td>
<td>Lam</td>
<td>ل (06B5), ل (06B6), ل (06B7)</td>
</tr>
<tr>
<td>م</td>
<td>0645</td>
<td>Meem</td>
<td>م (06FE)</td>
</tr>
<tr>
<td>ن</td>
<td>0646</td>
<td>Noon</td>
<td>ن (06B9), ن (06BA), ن (06BC)</td>
</tr>
<tr>
<td>ه</td>
<td>0647</td>
<td>Heh</td>
<td>ه (06BE), ه (06C1), ه (06D5)</td>
</tr>
<tr>
<td>و</td>
<td>0648</td>
<td>Waw</td>
<td>و (06C4), و (06C5), و (06C6), و (06C7), و (06C8)</td>
</tr>
<tr>
<td>ي</td>
<td>0649</td>
<td>Alef Maksura</td>
<td>ي (06CC), ي (06CD), ي (06CE)</td>
</tr>
<tr>
<td>ي</td>
<td>064A</td>
<td>Yeh</td>
<td>ب (067B), ب (067E), ي (06D0), ي (06D1), ي (06D2)</td>
</tr>
</tbody>
</table>
Issues Need Further Investigations
Similar Shape Characters

بابن

بابن

بيتي

بيتي

input[0] = U+0628
input[1] = U+067b
input[2] = U+067a
input[3] = U+06d0

input[0] = U+0628
input[1] = U+064a
input[2] = U+067a
input[3] = U+06d0

input[0] = U+0628
input[1] = U+064a
input[2] = U+062a
input[3] = U+06d0

input[0] = U+0628
input[1] = U+064a
input[2] = U+062a
input[3] = U+064a
Issues Need Further Investigations
Similar Shape Characters

input[0] = U+06be
input[1] = U+0631
input[2] = U+06a7
input[3] = U+06d0

input[0] = U+0647
input[1] = U+0631
input[2] = U+0641
input[3] = U+06d0

input[0] = U+0647
input[1] = U+0631
input[2] = U+0641
input[3] = U+064a

www.arabic-domains.org
Issues Need Further Investigations
Similar Shape Characters

إشارة

issues

input[0] = U+0673
input[1] = U+0634
input[2] = U+0627
input[3] = U+0631
input[4] = U+06c3

إشارة

issues

input[0] = U+0625
input[1] = U+0634
input[2] = U+0627
input[3] = U+0631
input[4] = U+06c3

إشارة

issues

input[0] = U+0625
input[1] = U+0634
input[2] = U+0627
input[3] = U+0631
input[4] = U+0629

www.arabic-domains.org
Issues Need Further Investigations
Similar Shape Characters

کلی
کلی
کلی

input[0] = U+06a9
input[1] = U+0644
input[2] = U+06cd

input[0] = U+06a9
input[1] = U+0644
input[2] = U+06cc

input[0] = U+0643
input[1] = U+0644
input[2] = U+0649
1

Arabic Label (U-Label): كلي
Unicode (U+): U+06CC, U+0644, U+06A9
ASCII Label (A-Label): xn--ghb5rwd.com

Now see where this domain will go: كلي.com

Why the domain name was registered:
1. To illustrate that an IDN script-based registration (particularly with Arabic Script) is a good ground for phishing.
2. To encourage the international communities in general and language communities (that use Arabic script) in particular to work together in finding solutions to these kind of problems.
- Arabic script domains will include characters that are LTR (e.g., dash, dot, AE digits)
- Exiting IDNA2003 (stringprep) assumes that the first and last characters of the label must be RandALCAt characters.
<table>
<thead>
<tr>
<th>Language VS. Script</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language</strong></td>
</tr>
<tr>
<td>Protocol</td>
</tr>
<tr>
<td>• Protocol works with scripts except for certain special cases</td>
</tr>
<tr>
<td>Registry</td>
</tr>
<tr>
<td>• Accepted character set table</td>
</tr>
<tr>
<td>• Guideline rules specific for language</td>
</tr>
<tr>
<td>• e.g., character folding</td>
</tr>
<tr>
<td>• Prohibition of mixing languages</td>
</tr>
</tbody>
</table>
Part IV: Conclusions

www.arabic-domains.org
Conclusions

- There are some work needed at the level of IDNA protocol
  - The current PVALID/DISALLOWED list should be updated
  - It should be for the whole script and suit all the languages
- Speedup the deployment of the fast track (At the TLD/Registry level)
  - Start with IDN ccTLDs and each one dedicated for only one language (not script)
    - Most ccTLDs are trying to help their local communities/language groups in a controllable manner that would help to reach a simple, comprehensive and more proper solutions to the users
  - Each language using Arabic script should be treated separately – a character set table per language
    - Standardize accepted character set tables per language that would be used for all IDN TLD registries for the same language

www.arabic-domains.org
Conclusions

- Scrip-based IDN TLD
  - A single script IDN TLD (ccTLD or gTLD) with multiple languages is not yet ready!
    - the current experience accumulated during the last few years is not enough to clear the path for allowing language mixing under a single TLD.
    - It needs to be studied very carefully and develop a universal character set table
  - Devolving a unified accepted character set table for the whole script is a challenging task
    - Needs cooperation among international bodies as well as language communities that are using the Arabic script

www.arabic-domains.org
Thank you

xn--mgbti28b
input[0] = U+0634
input[1] = U+06a9
input[2] = U+0631
input[3] = U+0627

xn--mgbti4d
input[0] = U+0634
input[1] = U+0643
input[2] = U+0631
input[3] = U+0627

www.arabic-domains.org